1272435

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

January 11, 2005

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.

APPLICATION NUMBER: ***60/529,432***
FILING DATE: ***December 12, 2003***
RELATED PCT APPLICATION NUMBER: ***PCT/US04/41883***

UNITED STATES PATENT AND TRADEMARK OFFICE
DEPARTMENT OF COMMERCE

Certified By

Jon W Dudas

Under Secretary
of Commerce for Intellectual Property
and Acting Director of the
Unites States Patent and Trademark Office

ORIGINAL

17621 U.S. PTO

PTO/SB/16 (05-03)
Approved for use through 4/30/2003. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No. EL844820434US **Date of Deposit: Dec. 12, 2003**

19587 U.S. PTO 60/529432 121203

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Nicole	Pauloski	Branford, CT
Li	Liu	West Haven, CT

Additional inventors are being named on the __________ *separately numbered sheets attached hereto*

TITLE OF THE INVENTION (500 characters max)

GENE EXPRESSION PROFILES AND METHODS OF USE

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

[x] Customer Number: 35969 → Place Customer Number Bar Code Label here: 35969 PATENT TRADEMARK OFFICE

Type Customer Number here

OR

[ ] Firm or Individual Name	
Address	
Address	
City	State: ZIP:
Country	Telephone: Fax:

ENCLOSED APPLICATION PARTS *(check all that apply)*

[x] Specification *Number of Pages* 30
[ ] CD(s), Number
[ ] Drawing(s) *Number of Sheets*
[x] Other (specify) **Return Receipt Postcard.**
[ ] Application Data Sheet. See 37 CFR 1.76

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

[ ] Applicant claims small entity status. See 37 CFR 1.27.
[ ] A check or money order is enclosed to cover the filing fees.
[x] The Director is hereby authorized to charge filing fees or credit any overpayment to Deposit Account Number: 13-3372
[ ] Payment by credit card. Form PTO-2038 is attached.

FILING FEE AMOUNT ($): 160.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

[x] No.
[ ] Yes, the name of the U.S. Government agency and the Government contract number are: __________

Respectfully submitted,

SIGNATURE Susan M. Pellegrino

Date 12/12/03

TYPED or PRINTED NAME Susan M. Pellegrino

REGISTRATION NO. *(if appropriate)* 48,972

Docket Number: 5185-Prov

TELEPHONE (203) 812-6450

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

This collection of information is required by 37 CFR 1.51. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 8 hours to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: **Mail Stop Provisional Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

GENE EXPRESSION PROFILES AND METHODS OF USE

FIELD OF THE INVENTION

[001] The present invention relates to gene expression profiles, microarrays comprising nucleic acid sequences representing gene expression profiles, and methods of using gene expression profiles and microarrays.

BACKGROUND OF THE INVENTION

[002] Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription of particular genes (e.g., through control of initiation, provision of RNA precursors, RNA processing, etc.). For example, losses and gains of genetic material play an important role in malignant transformation and progression. These gains and losses are thought to be "driven" by at least two kinds of genes, oncogenes and tumor suppressor genes. Oncogenes are positive regulators of tumorgenesis, while tumor suppressor genes are negative regulators of tumorgenesis (Marshall, Cell 64:313-326, 1991; Weinberg, Science 254:1138-1146, 1991). Therefore, one mechanism of activating unregulated growth is to increase the number of genes coding for oncogene proteins or to increase the level of expression of these oncogenes (e.g., in response to cellular or environmental changes), and another mechanism is to lose genetic material or to decrease the level of expression of genes that code for tumor suppressors. This model is supported by the losses and gains of genetic material associated with glioma progression (Mikkelson et al., J. Cellular Biochem. 46:3-8, 1991). Thus, changes in the expression (transcription) levels of particular genes (e.g., oncogenes or tumor suppressors) serve as signposts for the presence and progression of various cancers.

[003] Compounds which are used as therapeutics to treat these various diseases (e.g., cancer) presumably reverse some, or all, of these gene expression changes. The expression change of at least some of these genes may therefore, be used as a method to monitor, or even predict, the efficacy of such therapeutics. The analysis of these expression changes may be performed in the target tissue of interest (e.g., tumor) or in some surrogate cell population (e.g., peripheral blood leukocytes). In the latter case, correlation of the gene expression changes with efficacy (e.g., tumor shrinkage or non-growth) must be especially strong for the expression change pattern to be used as a marker for efficacy.

[004] Histone deacetylases (HDACs) are key enzymes in the regulation of gene expression via their effects on nuclear chromatin. HDACs remove acetyl groups from lysine residues of histones and other transcriptional regulators, reversing the effects of histone acetyltransferases (HATs). Unregulated HDAC activity has been linked to cancer and HDAC inhibitors have been shown to inhibit growth of human tumor cells in vitro (Donadelli, et al., Mol. Carcinog. 38: 59-69, 2003).

HDAC inhibitors have been shown to cause cell cycle arrest, differentiation, and/or apoptosis of many tumors (Yoshida, et al., Curr. Med. Chem. 10: 2351-2358, 2003). Several HDAC inhibitors have also demonstrated potent anit-tumor activity in human xenograft models (Arts, et al., Curr. Med. Chem. 10: 2343-2350, 2003). Therefore, the data suggests that small molecule inhibitors of HDAC activity will be an important therapeutic mechanism in the treatment of cancer.

SUMMARY OF THE INVENTION

[005] The present invention is directed to gene expression profiles, microarrays comprising nucleic acid sequences representing said gene expression profiles, and methods of using said gene expression profiles and microarrays.

[006] In a preferred embodiment of the present invention, the gene expression profile is an expression profile comprising one or more genes that demonstrate altered expression following exposure to a drug. In a particular embodiment, the expression profile comprises one or more genes that demonstrate altered expression following exposure to a histone deacetylase inhibitor.

[007] In another embodiment of the present invention, the gene expression profile may be an expression profile comprising one or more genes selected from the group consisting of the genes listed in Table 1. In a further embodiment of the present invention, the gene expression profiles comprise one or more biomarkers isolated from the group comprising the genes listed in Table 1.

[008] The present invention is also directed to the discovery of the gene expression profile of cancer cells. In particular, the gene expression profiles of the present invention represents the profiles of cancer cells following treatment with a histone deacetylase inhibitor. As described in the Examples and in Table 1, compound-treated cancer cells have genes which are expressed at higher levels (i.e., which are up-regulated) and genes which are expressed at lower levels (i.e., which are down-regulated) relative to cells of the same type in untreated subjects. In particular, as described in Table 1, it has been shown that following treatment with a histone deacetylase inhibitor, several genes are up-regulated or down-regulated in the cancer cells relative to the corresponding untreated cancer cells. Genes which are up-regulated or down-regulated are referred to herein as "genes characteristic of small molecule efficacy."

[009] Also within the scope of the present invention are microarrays comprising one or more genes that demonstrate altered expression following exposure to a histone deacetylase inhibitor. In another embodiment of the present invention, the microarray may be a microarray comprising one or more genes selected from the group consisting of the genes listed in Table 1. In a further embodiment, the microarray may be a microarray comprising one or more biomarkers isolated from the group comprising the genes listed in Table 1.

[010] This invention also relates to methods for using said microarrays which include, but are not limited to, screening the effects of a drug, for example, a histone deacetylase inhibitor, on tissue or cell samples, screening toxicity effects on tissue or cell samples, identifying a disease state in a tissue or cell sample, providing a patient diagnosis, predicting a patient's response to treatment, distinguishing between control and drug-treated samples, distinguishing between normal and tumor samples, discovering novel drugs, and determining the level of gene expression in a tissue or cell sample.

[011] Another embodiment of the present invention is a method for screening the effects of a drug, for example, a histone deacetylase inhibitor, on a tissue or cell sample comprising the step of analyzing the level of expression of one or more genes selected from the gene expression profiles, wherein the gene expression levels of the tissue or cell sample are analyzed before and after exposure to the drug, and a variation in the expression level of the gene is indicative of a drug effect.

[012] Another aspect of the present invention is a method for distinguishing between normal and disease states comprising the step of analyzing the level of expression of one or more genes selected from the gene expression profiles, wherein the gene expression levels of normal and disease tissues are analyzed, and a variation in the expression level of the gene is indicative of a disease state.

[013] The instant invention also provides a method for providing a patient diagnosis comprising the step of analyzing the level of expression of one or more genes selected from the gene expression profiles, wherein the gene expression levels of normal and patient samples are analyzed, and a variation in the expression level of a gene in the patient sample is diagnostic of a disease. The patient samples include, but are not limited to, blood, amniotic fluid, plasma, semen, bone marrow, and tissue biopsy.

[014] Another aspect of the present invention is a method for discovering novel drugs comprising the step of analyzing the level of expression of one or more genes selected from the gene expression profiles, wherein the gene expression levels of the cells are analyzed before and after exposure to the drug, and a variation in the expression level of the gene is indicative of drug efficacy.

[015] This invention is also related to methods of identifying biomarkers comprising the steps of selecting a set of biomarker genes from a gene expression profile representing a disease or drug treatment

DETAILED DESCRIPTION OF THE INVENTION

[016] It is to be understood that this invention is not limited to the particular methodology, protocols, cell lines, animal species or genera, constructs, and reagents described and as such may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims.

[017] It must be noted that as used herein and in the appended claims, the singular forms "a," "and," and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, reference to "a gene" is a reference to one or more genes and includes equivalents thereof known to those skilled in the art, and so forth.

[018] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs. Although any methods, devices, and materials similar or equivalent to those described herein can be used in the practice or testing of the invention, the preferred methods, devices and materials are now described.

[019] All publications and patents mentioned herein are hereby incorporated herein by reference for the purpose of describing and disclosing, for example, the constructs and methodologies that are described in the publications which might be used in connection with the presently described invention. The publications discussed above and throughout the text are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention.

Definitions

[020] For convenience, the meaning of certain terms and phrases employed in the specification, examples, and appended claims are provided below.

[021] The phrase "a corresponding normal cell of" or "normal cell corresponding to" or "normal counterpart cell of" a diseased cell refers to a normal cell of the same type as that of the diseased cell. For example, a corresponding normal cell of a B lymphoma cell is a B cell.

[022] An "address" on an array (e.g., a microarray) refers to a location at which an element, for example, an oligonucleotide, is attached to the solid surface of the array.

[023] The term "agonist," as used herein, is meant to refer to an agent that mimics or up-regulates (e.g., potentiates or supplements) the bioactivity of a protein. An agonist may be a wild-type protein or derivative thereof having at least one bioactivity of the wild-type protein. An agonist may also be a compound that up-regulates expression of a gene or which increases at least one

bioactivity of a protein. An agonist can also be a compound which increases the interaction of a polypeptide with another molecule, for example, a target peptide or nucleic acid.

[024] "Amplification," as used herein, relates to the production of additional copies of a nucleic acid sequence. For example, amplification may be carried out using polymerase chain reaction (PCR) technologies which are well known in the art. (*see, e.g.,* Dieffenbach, C. W. and G. S. Dveksler (1995) PCR Primer, A Laboratory Manual, Cold Spring Harbor Press, Plainview, N.Y.)

[025] "Antagonist," as used herein, is meant to refer to an agent that down-regulates (e.g., suppresses or inhibits) at least one bioactivity of a protein. For example, a histone deacetylase inhibitor is an example of such an antagonist. An antagonist may be a compound which inhibits or decreases the interaction between a protein and another molecule, for example, a target peptide or enzyme substrate. An antagonist may also be a compound that down-regulates expression of a gene or which reduces the amount of expressed protein present.

[026] The term "antibody," as used herein, is intended to include whole antibodies, for example, of any isotype (IgG, IgA, IgM, IgE, etc.), and includes fragments thereof which are also specifically reactive with a vertebrate (e.g., mammalian) protein. Antibodies may be fragmented using conventional techniques and the fragments screened for utility in the same manner as described above for whole antibodies. Thus, the term includes segments of proteolytically-cleaved or recombinantly-prepared portions of an antibody molecule that are capable of selectively reacting with a certain protein. Non-limiting examples of such proteolytic and/or recombinant fragments include Fab, F(ab')2, Fab', Fv, and single chain antibodies (scFv) containing a V[L] and/or V[H] domain joined by a peptide linker. The scFv's may be covalently or non-covalently linked to form antibodies having two or more binding sites. The subject invention includes polyclonal, monoclonal, or other purified preparations of antibodies and recombinant antibodies.

[027] The terms "array" or "matrix" refer to an arrangement of addressable locations or "addresses" on a device. The locations can be arranged in two-dimensional arrays, three-dimensional arrays, or other matrix formats. The number of locations may range from several to at least hundreds of thousands. Most importantly, each location represents a totally independent reaction site. A "nucleic acid array" refers to an array containing nucleic acid probes, such as oligonucleotides or larger portions of genes. The nucleic acid on the array is preferably single-stranded. Arrays wherein the probes are oligonucleotides are referred to as "oligonucleotide arrays" or "oligonucleotide chips." A "microarray," also referred to herein as a "biochip" or "biological chip," is an array of regions having a density of discrete regions of at least about $100/cm^2$, and preferably at least about $1000/cm^2$. The regions in a microarray have typical dimensions, for example, diameters, in the range of between about 10-250 μm, and are separated from other regions in the array by about the same distance.

[028] The term "biological sample", as used herein, refers to a sample obtained from an organism or from components (e.g., cells) of an organism. The sample may be of any biological tissue or fluid. The sample may be a "clinical sample" which is a sample derived from a patient. Such samples include, but are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

[029] The term "biomarker" refers to a gene or gene product which is up- or down-regulated in a compound-treated, diseased cell of a subject having the disease relative to a counterpart untreated diseased cell. That is, the gene or gene product is sufficiently specific to the treated cell that it may be used, optionally with other genes or gene products, to identify, predict or detect efficacy of a small molecule for the disease. Generally, a biomarker is a gene or gene product that is characteristic of efficacy of a compound in a diseased cell or the response of that diseased cell to treatment by the compound.

[030] A nucleotide sequence is "complementary" to another nucleotide sequence if each of the bases of the two sequences match, that is, are capable of forming Watson-Crick base pairs. The term "complementary strand" is used herein interchangeably with the term "complement." The complement of a nucleic acid strand may be the complement of a coding strand or the complement of a non-coding strand.

[031] "Detection agents of genes" refers to agents that can be used to specifically detect the gene or other biological molecules relating to it, for example, RNA transcribed from the gene or polypeptides encoded by the gene. Exemplary detection agents are nucleic acid probes, which hybridize to nucleic acids corresponding to the gene, and antibodies.

[032] "Differential gene expression pattern" between cell A and cell B refers to a pattern reflecting the differences in gene expression between cell A and cell B. A differential gene expression pattern may also be obtained between a cell at one time point and a cell at another time point, or between a cell incubated or contacted with a compound and a cell that has not been incubated with or contacted with the compound.

[033] The term "cancer" includes, but is not limited to, solid tumors, such as cancers of the breast, respiratory tract, brain, reproductive organs, digestive tract, urinary tract, eye, liver, skin, head and neck, thyroid, parathyroid, and their distant metastases. The term also includes lymphomas, sarcomas, and leukemias.

[034] Examples of breast cancer include, but are not limited to, invasive ductal carcinoma, invasive lobular carcinoma, ductal carcinoma *in situ*, and lobular carcinoma *in situ*.

[035] Examples of cancers of the respiratory tract include, but are not limited to, small-cell and non-small-cell lung carcinoma, as well as bronchial adenoma and pleuropulmonary blastoma.

[036] Examples of brain cancers include, but are not limited to, brain stem and hypophtalmic glioma, cerebellar and cerebral astrocytoma, medulloblastoma, ependymoma, as well as neuroectodermal and pineal tumor.

[037] Tumors of the male reproductive organs include, but are not limited to, prostate and testicular cancer. Tumors of the female reproductive organs include, but are not limited to, endometrial, cervical, ovarian, vaginal, and vulvar cancer, as well as sarcoma of the uterus.

[038] Tumors of the digestive tract include, but are not limited to, anal, colon, colorectal, esophageal, gallbladder, gastric, pancreatic, rectal, small-intestine, and salivary gland cancers.

[039] Tumors of the urinary tract include, but are not limited to, bladder, penile, kidney, renal pelvis, ureter, and urethral cancers.

[040] Eye cancers include, but are not limited to, intraocular melanoma and retinoblastoma.

[041] Examples of liver cancers include, but are not limited to, hepatocellular carcinoma (liver cell carcinomas with or without fibrolamellar variant), cholangiocarcinoma (intrahepatic bile duct carcinoma), and mixed hepatocellular cholangiocarcinoma.

[042] Skin cancers include, but are not limited to, squamous cell carcinoma, Kaposi's sarcoma, malignant melanoma, Merkel cell skin cancer, and non-melanoma skin cancer.

[043] Head-and-neck cancers include, but are not limited to, laryngeal / hypopharyngeal / nasopharyngeal / oropharyngeal cancer, and lip and oral cavity cancer.

[044] Lymphomas include, but are not limited to, AIDS-related lymphoma, non-Hodgkin's lymphoma, cutaneous T-cell lymphoma, Hodgkin's disease, and lymphoma of the central nervous system.

[045] Sarcomas include, but are not limited to, sarcoma of the soft tissue, osteosarcoma, malignant fibrous histiocytoma, lymphosarcoma, and rhabdomyosarcoma.

[046] Leukemias include, but are not limited to, acute myeloid leukemia, acute lymphoblastic leukemia, chronic lymphocytic leukemia, chronic myelogenous leukemia, and hairy cell leukemia.

[047] "A diseased cell of cancer" refers to a cell present in subjects having cancer. That is, a cell which is a modified form of a normal cell and is not present in a subject not having cancer, or a cell which is present in significantly higher or lower numbers in subjects having cancer relative to subjects not having cancer.

[048] The term "equivalent" is understood to include nucleotide sequences encoding functionally equivalent polypeptides. Equivalent nucleotide sequences may include sequences that differ by one or more nucleotide substitutions, additions, or deletions, such as allelic variants; and may, therefore, include sequences that differ from the nucleotide sequence of the nucleic acids referred to in Table 1 due to the degeneracy of the genetic code.

[049] The term "essentially all the genes of Table 1" refers to at least 90%, preferably at least 95% and most preferably at least 98% of the genes in Table 1.

[050] The term "expression profile," which is used interchangeably herein with "gene expression profile" and "fingerprint" of a cell refers to a set of values representing mRNA levels of one or more genes in a cell. An expression profile preferably comprises values representing expression levels of at least about 10 genes, preferably at least about 50, 100, 200 or more genes. Expression profiles may also comprise an mRNA level of a gene which is expressed at similar levels in multiple cells and conditions (e.g., a housekeeping gene such as GAPDH). For example, an expression profile of a diseased cell of cancer refers to a set of values representing mRNA levels of 10 or more genes in a diseased cell.

[051] The term "gene" refers to a nucleic acid sequence that comprises control and coding sequences necessary for the production of a polypeptide or precursor. The polypeptide can be encoded by a full length coding sequence or by any portion of the coding sequence. The gene may be derived in whole or in part from any source known to the art, including a plant, a fungus, an animal, a bacterial genome or episome, eukaryotic, nuclear or plasmid DNA, cDNA, viral DNA, or chemically synthesized DNA. A gene may contain one or more modifications in either the coding or the untranslated regions which could affect the biological activity or the chemical structure of the expression product, the rate of expression, or the manner of expression control. Such modifications include, but are not limited to, mutations, insertions, deletions, and substitutions of one or more nucleotides. The gene may constitute an uninterrupted coding sequence or it may include one or more introns, bound by the appropriate splice junctions.

[052] "Hybridization" refers to any process by which a strand of nucleic acid binds with a complementary strand through base pairing. For example, two single-stranded nucleic acids "hybridize" when they form a double-stranded duplex. The region of double-strandedness may include the full-length of one or both of the single-stranded nucleic acids, or all of one single-stranded nucleic acid and a subsequence of the other single-stranded nucleic acid, or the region of double-strandedness may include a subsequence of each nucleic acid. Hybridization also includes the formation of duplexes which contain certain mismatches, provided that the two strands are still forming a double-stranded helix. "Stringent hybridization conditions" refers to hybridization conditions resulting in essentially specific hybridization.

[053] The term "inhibitor," as used herein, is meant to refer to an agent that decreases (e.g., reduces or diminishes) the bioactivity of a protein. An inhibitor may also be a compound that alters expression of a gene or which decreases at least one bioactivity of a protein. An inhibitor can also be a compound which reduces or impedes the interaction of a polypeptide with another molecule, for example, a target peptide or nucleic acid.

[054] The term "isolated," as used herein, with respect to nucleic acids, such as DNA or RNA, refers to molecules separated from other DNAs or RNAs, respectively, that are present in the natural source of the macromolecule. The term "isolated" as used herein also refers to a nucleic acid or peptide that is substantially free of cellular material, viral material, culture medium when produced by recombinant DNA techniques, or chemical precursors or other chemicals when chemically synthesized. Moreover, an "isolated nucleic acid" may include nucleic acid fragments which are not naturally occurring as fragments and would not be found in the natural state. The term "isolated" is also used herein to refer to polypeptides which are isolated from other cellular proteins and is meant to encompass both purified and recombinant polypeptides.

[055] As used herein, the terms "label" and "detectable label" refer to a molecule capable of detection, including, but not limited to, radioactive isotopes, fluorophores, chemiluminescent moieties, enzymes, enzyme substrates, enzyme cofactors, enzyme inhibitors, dyes, metal ions, ligands (e.g., biotin or haptens), and the like. The term "fluorescer" refers to a substance or a portion thereof which is capable of exhibiting fluorescence in the detectable range. Particular examples of labels which may be used in the present invention include fluorescein, rhodamine, dansyl, umbelliferone, Texas red, luminol, NADPH, alpha-beta–galactosidase, and horseradish peroxidase.

[056] The phrase "level of expression of a gene in a cell" refers to the level of mRNA, as well as pre-mRNA nascent transcript(s), transcript processing intermediates, mature mRNA(s), and degradation products, encoded by a gene in the cell.

[057] As used herein, the term "nucleic acid" refers to polynucleotides such as deoxyribonucleic acid (DNA) and, where appropriate, ribonucleic acid (RNA). The term should also be understood to include, as equivalents, analogs of either RNA or DNA made from nucleotide analogs and, as applicable to the embodiment being described, single-stranded (sense or antisense) and double-stranded polynucleotides. Chromosomes, cDNAs, mRNAs, rRNAs, and ESTs are representative examples of molecules that may be referred to as nucleic acids.

[058] The phrase "nucleic acid corresponding to a gene" refers to a nucleic acid that can be used for detecting the gene, for example, a nucleic acid which is capable of hybridizing specifically to the gene.

[059] The phrase "nucleic acid sample derived from RNA" refers to one or more nucleic acid molecules (e.g., RNA or DNA) that may be synthesized from the RNA, and includes DNA produced from methods using PCR (e.g., RT-PCR).

[060] The term "oligonucleotide" as used herein refers to a nucleic acid molecule comprising, for example, from about 10 to about 1000 nucleotides. Oligonucleotides for use in the present invention are preferably from about 15 to about 150 nucleotides, more preferably from about 150 to about 1000 in length. The oligonucleotide may be a naturally occurring oligonucleotide or a synthetic oligonucleotide. Oligonucleotides may be prepared by the phosphoramidite method (Beaucage and Carruthers, Tetrahedron Lett. 22:1859-62, 1981), or by the triester method (Matteucci, et al., J. Am. Chem. Soc. 103:3185, 1981), or by other chemical methods known in the art.

[061] The term "percent identical" refers to sequence identity between two amino acid sequences or between two nucleotide sequences. For example, identity between two sequences may be determined by comparing a particular position in each sequence which may be aligned for purposes of comparison. When an equivalent position in the compared sequences is occupied by the same base or amino acid, then the molecules are identical at that position. When the equivalent site is occupied by the same or a similar amino acid residue (e.g., similar in steric and/or electronic nature), then the molecules may be referred to as homologous (similar) at that position. Expression as a percentage of homology, similarity, or identity refers to a function of the number of identical or similar amino acids at positions shared by the compared sequences. Various alignment algorithms and/or programs may be used including, for example, FASTA, BLAST, or ENTREZ. FASTA and BLAST are available as a part of the GCG sequence analysis package (University of Wisconsin, Madison, Wis.), and may be used with, for example, default settings. ENTREZ is available through the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD. In one embodiment, the percent identity of two sequences may be determined by the GCG program with a gap weight of 1 (e.g., each amino acid gap is weighted as if it were a single amino acid or nucleotide mismatch between the two sequences). Other techniques for alignment are described in Methods in Enzymology (vol. 266: Computer Methods for Macromolecular Sequence Analysis (1996), ed. Doolittle, Academic Press, Inc., a division of Harcourt Brace & Co., San Diego, California, USA). Preferably, an alignment program that permits gaps in the sequence is utilized to align the sequences. For example, the Smith-Waterman is one type of algorithm that permits gaps in sequence alignments (*see, e.g.,* Meth. Mol. Biol. 70:173-187, 1997). Also, the GAP program using the Needleman and Wunsch alignment method may be utilized to align sequences. An alternative search strategy uses MPSRCH software, which runs on a MASPAR computer. MPSRCH uses a

Smith-Waterman algorithm to score sequences on a massively parallel computer. This approach improves the ability to detect distantly related matches, and is especially tolerant of small gaps and nucleotide sequence errors. Nucleic acid-encoded amino acid sequences may be used to search both protein and DNA databases. Databases with individual sequences are described in Methods in Enzymology, ed. Doolittle, *supra*. Databases include, for example, Genbank, EMBL, and DNA Database of Japan (DDBJ).

[062] "Perfectly matched" in reference to a duplex means that the polynucleotide or oligonucleotide strands of the duplex form a double-stranded structure with one another such that every nucleotide in each strand undergoes Watson-Crick base pairing with a nucleotide in the other strand. The term also encompasses the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. A mismatch in a duplex between a target polynucleotide and an oligonucleotide or polynucleotide means that a pair of nucleotides in the duplex fails to undergo Watson-Crick base pairing. In reference to a triplex, this term means that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a base pair of the perfectly matched duplex.

[063] As used herein, a nucleic acid or other molecule attached to an array is referred to as a "probe" or "capture probe." When an array contains several probes corresponding to one gene, these probes are referred to as a "gene-probe set." A gene-probe set may consist of, for example, about 2 to about 20 probes, preferably from about 2 to about 10 probes, and most preferably about 5 probes.

[064] The "profile" of a cell's biological state refers to the levels of various constituents of a cell that are known to change in response to drug treatments and other perturbations of the biological state of the cell. Constituents of a cell include, for example, levels of RNA, levels of protein abundances, or protein activity levels.

[065] The term "protein" is used interchangeably herein with the terms "peptide" and "polypeptide."

[066] An expression profile in one cell is "similar" to an expression profile in another cell when the level of expression of the genes in the two profiles are sufficiently similar that the similarity is indicative of a common characteristic, for example, the same type of cell. Accordingly, the expression profiles of a first cell and a second cell are similar when at least 75% of the genes that are expressed in the first cell are expressed in the second cell at a level that is within a factor of two relative to the first cell.

[067] "Small molecule," as used herein, refers to a composition with a molecular weight of less than about 5 kD and most preferably less than about 4 kD. Small molecules can be nucleic acids, peptides, polypeptides, peptidomimetics, carbohydrates, lipids, or other organic or inorganic molecules. Many pharmaceutical companies have extensive libraries of chemical and/or biological mixtures, often fungal, bacterial, or algal extracts, which can be screened with any of the assays of the invention to identify compounds that modulate a bioactivity.

[068] The term "specific hybridization" of a probe to a target site of a template nucleic acid refers to hybridization of the probe predominantly to the target, such that the hybridization signal can be clearly interpreted. As further described herein, such conditions resulting in specific hybridization vary depending on the length of the region of homology, the GC content of the region, and the melting temperature ("Tm") of the hybrid. Thus, hybridization conditions may vary in salt content, acidity, and temperature of the hybridization solution and the washes.

[069] The phrase "value representing the level of expression of a gene" refers to a raw number which reflects the mRNA level of a particular gene in a cell or biological sample, for example, obtained from experiments for measuring RNA levels.

[070] A "variant" of polypeptide refers to a polypeptide having an amino acid sequence in which one or more amino acid residues is altered. The variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties (e.g., replacement of leucine with isoleucine). A variant may also have "nonconservative" changes (e.g., replacement of glycine with tryptophan). Analogous minor variations may include amino acid deletions or insertions, or both. Guidance in determining which amino acid residues may be substituted, inserted, or deleted without abolishing biological or immunological activity may be identified using computer programs well known in the art, for example, LASERGENE software (DNASTAR).

[071] The term "variant," when used in the context of a polynucleotide sequence, may encompass a polynucleotide sequence related to that of a particular gene or the coding sequence thereof. This definition may also include, for example, "allelic," "splice," "species," or "polymorphic" variants. A splice variant may have significant identity to a reference molecule, but will generally have a greater or lesser number of polynucleotides due to alternate splicing of exons during mRNA processing. The corresponding polypeptide may possess additional functional domains or an absence of domains. Species variants are polynucleotide sequences that vary from one species to another. The resulting polypeptides generally will have significant amino acid identity relative to each other. A polymorphic variant is a variation in the polynucleotide sequence of a particular gene between individuals of a given species. Polymorphic variants also may encompass "single nucleotide polymorphisms" (SNPs) in which the polynucleotide sequence varies by one base. The

presence of SNPs may be indicative of, for example, a certain population, a disease state, or a propensity for a disease state.

Microarrays for Determining the Level of Expression of Genes Characteristic of Small Molecule Efficacy

[072] Generally, determining expression profiles with microarrays involves the following steps: (a) obtaining an mRNA sample from a subject and preparing labeled nucleic acids therefrom (the "target nucleic acids" or "targets"); (b) contacting the target nucleic acids with an array under conditions sufficient for the target nucleic acids to bind to the corresponding probes on the array, for example, by hybridization or specific binding; (c) optional removal of unbound targets from the array; (d) detecting the bound targets, and (e) analyzing the results, for example, using computer based analysis methods. As used herein, "nucleic acid probes" or "probes" are nucleic acids attached to the array, whereas "target nucleic acids" are nucleic acids that are hybridized to the array. Each of these steps is described in more detail below.

Obtaining an mRNA sample of a subject

[073] Nucleic acid specimens may be obtained from an individual to be tested using either "invasive" or "non-invasive" sampling means. A sampling means is said to be "invasive" if it involves the collection of nucleic acids from within the skin or organs of an animal (including murine, human, ovine, equine, bovine, porcine, canine, or feline animal). Examples of invasive methods include blood collection, semen collection, needle biopsy, pleural aspiration, umbilical cord biopsy, etc. Examples of such methods are discussed by Kim et al., (J. Virol. 66:3879-3882, 1992); Biswas et al., (Ann. NY Acad. Sci. 590:582-583, 1990); and Biswas et al., (J. Clin. Microbiol. 29:2228-2233, 1991).

[074] In contrast, a "non-invasive" sampling means is one in which the nucleic acid molecules are recovered from an internal or external surface of the animal. Examples of such "non-invasive" sampling means include, for example, "swabbing," collection of tears, saliva, urine, fecal material, sweat or perspiration, hair, etc.

[075] In one embodiment of the present invention, one or more cells from the subject to be tested are obtained and RNA is isolated from the cells. In a preferred embodiment, a sample of peripheral blood leukocytes (PBLs) cells is obtained from the subject. It is also possible to obtain a cell sample from a subject, and then to enrich the sample for a desired cell type. For example, cells may be isolated from other cells using a variety of techniques, such as isolation with an antibody binding to an epitope on the cell surface of the desired cell type. Where the desired cells are in a solid tissue, particular cells may be dissected, for example, by microdissection or by laser capture microdissection (LCM) (*see, e.g.*, Bonner et al., Science 278:1481, 1997; Emmert-Buck et

al., Science 274:998, 1996; Fend et al., Am. J. Path. 154:61, 1999; and Murakami et al., Kidney Int. 58:1346, 2000).

[076] RNA may be extracted from tissue or cell samples by a variety of methods, for example, guanidium thiocyanate lysis followed by CsCl centrifugation (Chirgwin et al., Biochemistry 18:5294-5299, 1979). RNA from single cells may be obtained as described in methods for preparing cDNA libraries from single cells (*see, e.g.*, Dulac, Curr. Top. Dev. Biol. 36:245, 1998; Jena et al., J. Immunol. Methods 190:199, 1996).

[077] The RNA sample can be further enriched for a particular species. In one embodiment, for example, poly(A)+ RNA may be isolated from an RNA sample. In particular, poly-T oligonucleotides may be immobilized on a solid support to serve as affinity ligands for mRNA. Kits for this purpose are commercially available, for example, the MessageMaker kit (Life Technologies, Grand Island, NY).

[078] In a preferred embodiment, the RNA population may be enriched for sequences of interest, such as the genes characteristic of small molecule efficacy. Enrichment may be accomplished, for example, by primer-specific cDNA synthesis, or multiple rounds of linear amplification based on cDNA synthesis and template-directed *in vitro* transcription (*see, e.g.*, Wang et al., Proc. Natl. Acad. Sci. USA 86:9717, 1989; Dulac et al., *supra*; Jena et al., *supra*).

[079] The population of RNA, enriched or not in particular species or sequences, may be further amplified. Such amplification is particularly important when using RNA from a single cell or a few cells. A variety of amplification methods are suitable for use in the methods of the present invention, including, for example, PCR; ligase chain reaction (LCR) (*see, e.g.*, Wu and Wallace, Genomics 4:560, 1989; Landegren et al., Science 241:1077, 1988); self-sustained sequence replication (SSR) (*see, e.g.*, Guatelli et al., Proc. Natl. Acad. Sci. USA 87:1874, 1990); nucleic acid based sequence amplification (NASBA) and transcription amplification (*see, e.g.*, Kwoh et al., Proc. Natl. Acad. Sci. USA 86:1173, 1989). Methods for PCR technology are well known in the art (*see, e.g.*, PCR Technology: Principles and Applications for DNA Amplification (ed. H. A. Erlich, Freeman Press, N.Y., N.Y., 1992); PCR Protocols: A Guide to Methods and Applications (eds. Innis, et al., Academic Press, San Diego, Calif., 1990); Mattila et al., Nucleic Acids Res. 19:4967, 1991; Eckert et al., PCR Methods and Applications 1:17, 1991; PCR (eds. McPherson et al., IRL Press, Oxford); and U.S. Pat. No. 4,683,202). Methods of amplification are described, for example, by Ohyama et al., (BioTechniques 29:530, 2000); Luo et al., (Nat. Med. 5:117, 1999); Hegde et al., (BioTechniques 29:548, 2000); Kacharmina et al., (Meth. Enzymol. 303:3, 1999); Livesey et al., Curr. Biol. 10:301, 2000); Spirin et al., (Invest. Ophtalmol. Vis. Sci. 40:3108, 1999); and Sakai et al., (Anal. Biochem. 287:32, 2000). RNA amplification and cDNA synthesis

may also be conducted in cells *in situ* (*see, e.g.*, Eberwine et al. Proc. Natl. Acad. Sci. USA 89:3010, 1992).

Labeling of the nucleic acids to be analyzed

[080] Generally, the target molecules will be labeled to permit detection of hybridization of the target molecules to a microarray. That is, the probe may comprise a member of a signal producing system and thus, is detectable, either directly or through combined action with one or more additional members of a signal producing system. Examples of directly detectable labels include isotopic and fluorescent moieties incorporated, usually by a covalent bond, into a moiety of the probe, such as a nucleotide monomeric unit (e.g., dNMP of the primer), or a photoactive or chemically active derivative of a detectable label which can be bound to a functional moiety of the probe molecule.

[081] Nucleic acids may be labeled during or after enrichment and/or amplification of RNAs. For example, reverse transcription may be carried out in the presence of a dNTP conjugated to a detectable label, most preferably a fluorescently labeled dNTP. In another embodiment, the cDNA or RNA probe may be synthesized in the absence of detectable label and may be labeled subsequently, for example, by incorporating biotinylated dNTPs or rNTP, or some similar means (e.g., photo-cross-linking a psoralen derivative of biotin to RNAs), followed by addition of labeled streptavidin (e.g., phycoerythrin-conjugated streptavidin) or the equivalent.

[082] Fluorescent moieties or labels of interest include coumarin and its derivatives (e.g., 7-amino-4-methylcoumarin, aminocoumarin); bodipy dyes such as Bodipy FL and cascade blue; fluorescein and its derivatives (e.g., fluorescein isothiocyanate, Oregon green); rhodamine dyes (e.g., Texas red, tetramethylrhodamine); eosins and erythrosins; cyanine dyes (e.g., Cy2, Cy3, Cy3.5, Cy5, Cy5.5, Cy7); FluorX, macrocyclic chelates of lanthanide ions (e.g., quantum dye™); fluorescent energy transfer dyes such as thiazole orange-ethidium heterodimer, TOTAB, dansyl, etc. Individual fluorescent compounds which have functionalities for linking to an element desirably detected in an apparatus or assay of the invention, or which may be modified to incorporate such functionalities may also be utilized (*see, e.g.*, Kricka, 1992, Nonisotopic DNA Probe Techniques, Academic Press San Diego, Calif.).

[083] Chemiluminescent labels include luciferin and 2,3-dihydrophthalazinediones, for example, luminol.

[084] Labels may also be members of a signal producing system that act in concert with one or more additional members of the same system to provide a detectable signal. Illustrative of such labels are members of a specific binding pair, such as ligands, for example, biotin, fluorescein, digoxigenin, antigen, polyvalent cations, chelator groups and the like. Members may specifically

bind to additional members of the signal producing system, and the additional members may provide a detectable signal either directly or indirectly, for example, an antibody conjugated to a fluorescent moiety or an enzymatic moiety capable of converting a substrate to a chromogenic product (e.g., alkaline phosphatase conjugate antibody and the like).

[085] Additional labels of interest include those that provide a signal only when the probe with which it is associated is specifically bound to a target molecule. Such labels include "molecular beacons" as described in Tyagi and Kramer (Nature Biotech. 14:303, 1996) and EP 0 070 685 B1. Other labels of interest include those described in U.S. Patent No. 5,563,037; WO 97/17471; and WO 97/17076.

[086] In other embodiments, the target nucleic acid may not be labeled. In this case, hybridization may be determined, for example, by plasmon resonance (*see, e.g.*, Thiel et al. Anal. Chem. 69:4948, 1997).

[087] In one embodiment, a plurality (e.g., 2, 3, 4, 5, or more) of sets of target nucleic acids are labeled and used in one hybridization reaction ("multiplex" analysis). For example, one set of nucleic acids may correspond to RNA from one cell and another set of nucleic acids may correspond to RNA from another cell. The plurality of sets of nucleic acids may be labeled with different labels, for example, different fluorescent labels (e.g., fluorescein and rhodamine) which have distinct emission spectra so that they can be distinguished. The sets may then be mixed and hybridized simultaneously to one microarray (*see, e.g.,* Shena et al., Science 270:467-470, 1995).

[088] Examples of distinguishable labels for use when hybridizing a plurality of target nucleic acids to one array are well known in the art and include: two or more different emission wavelength fluorescent dyes such as Cy3 and Cy5; combination of fluorescent proteins and dyes such as phicoerythrin and Cy5; two or more isotopes with different energy of emission such as ^{32}P and ^{33}P; gold or silver particles with different scattering spectra; labels which generate signals under different treatment conditions such as temperature, pH, treatment with additional chemical agents, etc.; or generate signals at different time points after treatment. Using one or more enzymes for signal generation allows for the use of an even greater variety of distinguishable labels, based on different substrate specificity of enzymes (e.g., alkaline phosphatase/peroxidase).

[089] The quality of labeled nucleic acids may be evaluated prior to hybridization to an array. In one embodiment, the GeneChip® Test3 Array from Affymetrix (Santa Clara, CA) may be used for that purpose. This array contains probes representing a subset of characterized genes from several organisms including mammals. Thus, the quality of a labeled nucleic acid sample can be determined by hybridization of a fraction of the sample to an array.

Exemplary microarrays

[090] Preferred microarrays for use according to the invention include one or more probes of genes characteristic of small molecule efficacy. In a preferred embodiment, the microarray comprises probes corresponding to one or more of genes selected from the group consisting of genes which are up-regulated in cancer and genes which are down-regulated in cancer. The microarray may comprise probes corresponding to at least 10, preferably at least 20, at least 50, at least 100 or at least 1000 genes characteristic of small molecule efficacy. The microarray may comprise probes corresponding to each gene listed in Table 1.

[091] There may be one or more than one probe corresponding to each gene on a microarray. For example, a microarray may contain from 2 to 20 probes corresponding to one gene and preferably about 5 to 10. The probes may correspond to the full-length RNA sequence or complement thereof of genes characteristic of small molecule efficacy, or the probe may correspond to a portion thereof, which portion is of sufficient length to permit specific hybridization. Such probes may comprise from about 50 nucleotides to about 100, 200, 500, or 1000 nucleotides or more than 1000 nucleotides. As further described herein, microarrays may contain oligonucleotide probes, consisting of about 10 to 50 nucleotides, preferably about 15 to 30 nucleotides and more preferably about 20-25 nucleotides. The probes are preferably single-stranded and will have sufficient complementarity to its target to provide for the desired level of sequence specific hybridization.

[092] Typically, the arrays used in the present invention will have a site density of greater than 100 different probes per cm^2. Preferably, the arrays will have a site density of greater than $500/cm^2$, more preferably greater than about $1000/cm^2$, and most preferably, greater than about $10{,}000/cm^2$. Preferably, the arrays will have more than 100 different probes on a single substrate, more preferably greater than about 1000 different probes, still more preferably, greater than about 10,000 different probes and most preferably, greater than 100,000 different probes on a single substrate.

[093] A number of different microarray configurations and methods for their production are known to those of skill in the art and are disclosed in U.S. Patent Nos: 5,242,974; 5,384,261; 5,405,783; 5,412,087; 5,424,186; 5,429,807; 5,436,327; 5,445,934; 5,556,752; 5,405,783; 5,412,087; 5,424,186; 5,429,807; 5,436,327; 5,472,672; 5,527,681; 5,529,756; 5,545,531; 5,554,501; 5,561,071; 5,571,639; 5,593,839; 5,624,711; 5,700,637; 5,744,305; 5,770,456; 5,770,722; 5,837,832; 5,856,101; 5,874,219; 5,885,837; 5,919,523; 6,022,963; 6,077,674; and 6,156,501; Shena et al., Tibtech 16:301, 1998; Duggan et al., Nat. Genet. 21:10, 1999; Bowtell et al., Nat. Genet. 21:25, 1999; Lipshutz et al., 21 Nature Genet. 20-24, 1999; Blanchard et al., 11 Biosensors and Bioelectronics, 687-90, 1996; Maskos et al., 21 Nucleic Acids Res. 4663-69, 1993;

Hughes et al., Nat. Biotechol. (2001) 19:342; the disclosures of which are herein incorporated by reference. Patents describing methods of using arrays in various applications include: U.S. Pat. Nos. 5,143,854; 5,288,644; 5,324,633; 5,432,049; 5,470,710; 5,492,806; 5,503,980; 5,510,270; 5,525,464; 5,547,839; 5,580,732; 5,661,028; 5,848,659; and 5,874,219; the disclosures of which are herein incorporated by reference.

[094] Arrays preferably include control and reference nucleic acids. Control nucleic acids include, for example, prokaryotic genes such as bioB, bioC and bioD, cre from P1 bacteriophage or polyA controls, such as dap, lys, phe, thr, and trp. Reference nucleic acids allow the normalization of results from one experiment to another and the comparison of multiple experiments on a quantitative level. Exemplary reference nucleic acids include housekeeping genes of known expression levels, for example, GAPDH, hexokinase, and actin.

[095] In one embodiment, an array of oligonucleotides may be synthesized on a solid support. Exemplary solid supports include glass, plastics, polymers, metals, metalloids, ceramics, organics, etc. Using chip masking technologies and photoprotective chemistry, it is possible to generate ordered arrays of nucleic acid probes. These arrays, which are known, for example, as "DNA chips" or very large scale immobilized polymer arrays ("VLSIPS™" arrays), may include millions of defined probe regions on a substrate having an area of about 1 cm^2 to several cm^2, thereby incorporating from a few to millions of probes (*see, e.g.*, U.S. Patent No. 5,631,734).

[096] cDNA probes may be prepared according to methods known in the art and further described herein, for example, by reverse-transcription PCR (RT-PCR) of RNA using sequence specific primers. Oligonucleotide probes may also be synthesized chemically. Sequences of genes or cDNA from which probes are generated may be obtained, for example, from GenBank, other public databases, or publications.

[097] Nucleic acid probes may be natural nucleic acids or chemically modified nucleic acids (e.g., composed of nucleotide analogs); however, the probes should possess activated hydroxyl groups compatible with the linking chemistry. The protective groups may be photolabile, or the protective groups may be labile under certain chemical conditions (e.g., acid). The surface of the solid support may contain a composition that generates acids upon exposure to light. Thus, exposure of a region of the substrate to light generates acids in that region that remove the protective groups in the exposed region. Also, the synthesis method may use 3'- protected 5'-0-phosphoramidite-activated deoxynucleoside. In this case, the oligonucleotide is synthesized in the 5' to 3' direction, which results in a free 5' end.

[098] In one embodiment of the present invention, oligonucleotides of an array may be synthesized using a 96-well automated multiplex oligonucleotide synthesizer (A.M.O.S.) that is

capable of producing thousands of oligonucleotides (*see, e.g.*, Lashkari et al., Proc. Natl. Acad. Sci. USA 93: 7912, 1995).

Hybridization of the target nucleic acids to the microarray

[099] To compare expression levels, labeled nucleic acids may be contacted with the array under conditions sufficient for binding between the target nucleic acid and the probe on the array. In a preferred embodiment, the hybridization conditions may be selected to provide for the desired level of hybridization specificity; that is, conditions sufficient for hybridization to occur between the labeled nucleic acids and probes on the microarray.

[100] Hybridization may be carried out in conditions permitting essentially specific hybridization. The length and GC content of the nucleic acid will determine the thermal melting point and thus, the hybridization conditions necessary for obtaining specific hybridization of the probe to the target nucleic acid. These factors are well known to a person of skill in the art, and may also be tested in assays. An extensive guide to nucleic acid hybridization may be found in Tijssen et al. (Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993)). Generally, stringent conditions may be selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Highly stringent conditions may be selected to be equal to the Tm point for a particular probe. Sometimes the term "dissociation temperature" (Td) is used to define the temperature at which at least half of the probe dissociates from a perfectly matched target nucleic acid. In any case, a variety of techniques for estimating the Tm or Td are available, and generally are described in Tijssen, *supra*. Typically, G-C base pairs in a duplex are estimated to contribute about 3°C to the Tm, while A-T base pairs are estimated to contribute about 2°C, up to a theoretical maximum of about 80-100°C. However, more sophisticated models of Tm and Td are available in which G-C stacking interactions, solvent effects, the desired assay temperature, and the like are taken into account. For example, probes may be designed to have a dissociation temperature (Td) of approximately 60°C, using the formula:

$$Td = \frac{[((3 \times \#GC) + (2 \times \#AT)) \times 37] - 562}{\#bp} - 5;$$

where #GC, #AT, and #bp are the number of guanine-cytosine base pairs, the number of adenine-thymine base pairs, and the number of total base pairs, respectively, involved in the annealing of the probe to the template DNA.

[101] In a preferred embodiment, non-specific binding or background signal may be reduced by the use of a detergent (e.g, C-TAB) or a blocking reagent (e.g., sperm DNA, cot-1 DNA, etc.) during the hybridization. In a particularly preferred embodiment, the hybridization may be performed in the presence of about 0.5 mg/ml DNA (e.g., herring sperm DNA). The use of blocking agents in hybridization is well known to those of skill in the art (*see, e.g.,* Tijssen, *supra*).

[102] If the target sequences are detected using the same label, different arrays may be employed for each physiological source or the same array may be screened multiple times. The above methods may be varied to provide for multiplex analysis by employing different and distinguishable labels for the different target populations (e.g., different physiological sources). According to this multiplex method, the same array may be used at the same time for each of the different target populations.

Detection of hybridization and analysis of results

[103] The methods described above result in the production of hybridization patterns of labeled target nucleic acids on the array surface. The resultant hybridization patterns of labeled nucleic acids may be visualized or detected in a variety of ways, with the particular manner of detection selected based on the particular label of the target nucleic acid. Representative detection means include scintillation counting, autoradiography, fluorescence measurement, colorimetric measurement, light emission measurement, light scattering, and the like.

[104] One such method of detection utilizes an array scanner that is commercially available (Affymetrix, Santa Clara, CA), for example, the 417™ Arrayer, the 418™ Array Scanner, or the Agilent GeneArray™ Scanner. This scanner is controlled from a system computer with an interface and easy-to-use software tools. The output may be directly imported into or directly read by a variety of software applications. Preferred scanning devices are described in, for example, U.S. Patent Nos. 5,143,854 and 5,424,186.

[105] For fluorescent labeled probes, the fluorescence emissions at each site of a transcript array may be, preferably, detected by scanning confocal laser microscopy. Alternatively, a laser may be used that allows simultaneous specimen illumination at wavelengths specific to the two fluorophores and emissions from the two fluorophores may be analyzed simultaneously (*see, e.g.,* Shalon et al., Genome Res. 6:639-645, 1996). In a preferred embodiment, the arrays may be scanned with a laser fluorescent scanner with a computer controlled X-Y stage and a microscope objective. Fluorescence laser scanning devices are described in Shalon et al., *supra*.

[106] Following the data gathering operation, the data will typically be reported to a data analysis operation. To facilitate the sample analysis operation, the data obtained by the reader from the device may be analyzed using a digital computer. Typically, the computer will be appropriately

programmed for receipt and storage of the data from the device, as well as for analysis and reporting of the data gathered, for example, subtraction of the background, deconvolution of multi-color images, flagging or removing artifacts, verifying that controls have performed properly, normalizing the signals, interpreting fluorescence data to determine the amount of hybridized target, normalization of background and single base mismatch hybridizations, and the like.

[107] In a preferred embodiment, a system comprises a search function that allows one to search for specific patterns, for example, patterns relating to differential gene expression, for example, between the expression profile of a cancer cell and the expression profile of a counterpart normal cell in a subject. A system preferably allows one to search for patterns of gene expression between more than two samples.

[108] Various algorithms are available for analyzing gene expression profile data, for example, the type of comparisons to perform. In certain embodiments, it is desirable to group genes that are co-regulated. This allows for the comparison of large numbers of profiles. A preferred embodiment for identifying such groups of genes involves clustering algorithms (for reviews of clustering algorithms, *see, e.g.*, Fukunaga, 1990, Statistical Pattern Recognition, 2nd Ed., Academic Press, San Diego; Everitt, 1974, Cluster Analysis, London: Heinemann Educ. Books; Hartigan, 1975, Clustering Algorithms, New York: Wiley; Sneath and Sokal, 1973, Numerical Taxonomy, Freeman; Anderberg, 1973, Cluster Analysis for Applications, Academic Press: New York).

[109] Clustering may be based on other characteristics of the genes, for example, their level of expression (*see, e.g.*, U.S. Patent No. 6,203,987), or permit clustering of time curves (*see, e.g.* U.S. Patent No. 6,263,287). Examples of clustering algorithms include K-means clustering and hierarchical clustering. Clustering may also be achieved by visual inspection of gene expression data using a graphical representation of the data (e.g. a "heat map"). An example of software which contains clustering algorithms and a means to graphically represent gene expression data is Spotfire DecisionSite (Spotfire, Inc., Somerville, Massachusetts and Göteborg, Sweden).

Data Analysis Methods

[110] Comparison of the expression levels of one or more genes characteristic of small molecule efficacy with reference expression levels, for example, expression levels in diseased cells of cancer or in normal counterpart cells, is preferably conducted using computer systems. In one embodiment, expression levels may be obtained from two cells and these two sets of expression levels may be introduced into a computer system for comparison. In a preferred embodiment, one set of expression levels is entered into a computer system for comparison with values that are

already present in the computer system, or in computer-readable form that is then entered into the computer system.

[111] In a preferred embodiment, the computer system may also contain a database comprising values representing levels of expression of one or more genes characteristic of small molecule efficacy. The database may contain one or more expression profiles of genes characteristic of small molecule efficacy in different cells.

[112] In another embodiment, the invention provides a computer-readable form of the gene expression profile data, or of values corresponding to the level of expression of at least one gene characteristic of cancer in a diseased cell. The values may be mRNA expression levels obtained from experiments, for example, microarray analysis. The values may also be mRNA levels normalized relative to a reference gene whose expression is constant in numerous cells under numerous conditions (e.g., GAPDH). In other embodiments, the values in the computer may be ratios of, or differences between, normalized or non-normalized mRNA levels in different samples.

[113] In one embodiment, the expression profiles expression profiles from cancer cells of one or more subjects, which cells are treated *in vivo* or *in vitro* with a drug, for example, a histone deacetylase inhibitor used for therapy of cancer. Expression data of a cell of a subject treated *in vitro* or *in vivo* with the drug is entered into a computer and the computer is instructed to compare the data entered to the data in the computer, and to provide results indicating whether the expression data input into the computer are more similar to those of a cell of a subject that is responsive to the drug or more similar to those of a cell of a subject that is not responsive to the drug. Thus, the results indicate whether the subject is likely to respond to the treatment with the drug or unlikely to respond to it.

[114] The invention also provides a machine-readable or computer-readable medium including program instructions for performing the following steps: (i) comparing a plurality of values corresponding to expression levels of one or more genes characteristic of small molecule efficacy in a query cell with a database including records comprising reference expression or expression profile data of one or more reference cells and an annotation of the type of cell; and (ii) indicating to which cell the query cell is most similar based on similarities of expression profiles. The reference cells may be cells from subjects at different stages of cancer. The reference cells may also be cells from subjects responding or not responding to a particular drug treatment and optionally incubated *in vitro* or *in vivo* with the drug.

[115] The reference cells may also be cells from subjects responding or not responding to several different treatments, and the computer system indicates a preferred treatment for the subject.

Accordingly, the invention provides a method for selecting a therapy for a patient having cancer, the method comprising: (i) providing the level of expression of one or more genes characteristic of small molecule efficacy in a diseased cell of the patient; (ii) providing a plurality of reference profiles, each associated with a therapy, wherein the subject expression profile and each reference profile has a plurality of values, each value representing the level of expression of a gene characteristic of cancer; and (iii) selecting the reference profile most similar to the subject expression profile, to thereby select a therapy for said patient. In a preferred embodiment, step (iii) is performed by a computer. The most similar reference profile may be selected by weighing a comparison value of the plurality using a weight value associated with the corresponding expression data.

[116] The relative abundance of an mRNA in two biological samples may be scored as a perturbation and its magnitude determined (i.e., the abundance is different in the two sources of mRNA tested), or as not perturbed (i.e., the relative abundance is the same). In various embodiments, a difference between the two sources of RNA of at least a factor of about 25% (RNA from one source is 25% more abundant in one source than the other source), more usually about 50%, even more often by a factor of about 2 (twice as abundant), 3 (three times as abundant) or 5 (five times as abundant) is scored as a perturbation. Perturbations may be used by a computer for calculating and expression comparisons.

Exemplary Diagnostic and Prognostic Compositions and Devices

[117] In one embodiment, the invention provides a composition comprising a plurality of detection agents for detecting expression of genes characteristic of small molecule efficacy. In a preferred embodiment, the composition comprises at least 2, preferably at least 3, 5, 10, 20, 50, or 100 different detection agents. A detection agent may be a nucleic acid probe, for example, DNA or RNA, or it may be a polypeptide, for example, an antibody that binds to the polypeptide encoded by a gene characteristic of cancer. The probes may be present in equal amount or in different amounts in the solution.

[118] A nucleic acid probe may be at least about 10 nucleotides long, preferably at least about 15, 20, 25, 30, 50, 100 nucleotides or more, and may comprise the full-length gene. Preferred probes are those that hybridize specifically to the genes listed in Table 1.

[119] Nucleic acid probes may be obtained, for example, by polymerase chain reaction (PCR) amplification of gene segments from genomic DNA, cDNA (e.g., by RT-PCR), or cloned sequences. Sequences may also be obtained from GenBank or other public sources.

[120] Oligonucleotides of the invention may be synthesized by standard methods known in the art, for example, by an automated DNA synthesizer. As an example, phosphorothioate

oligonucleotides may be synthesized by the method of Stein et al., (Nucl. Acids Res. 16:3209, 1988), and methylphosphonate oligonucleotides may be prepared by controlled pore glass polymer supports (*see, e.g.,* Sarin et al., Proc. Nat. Acad. Sci. U.S.A. 85:7448-7451, 1988). In another embodiment, the oligonucleotide may be a 2'-0-methylribonucleotide (Inoue et al., Nucl. Acids Res. 15:6131-6148, 1987), or a chimeric RNA-DNA analog (Inoue et al., FEBS Lett. 215:327-330, 1987).

[121] Probes of the gene sequences listed in Table 1 may also be generated synthetically. Single-step assembly of a gene from large numbers of oligodeoxyribonucleotides may be accomplished as described by Stemmer et al., (Gene 164:49-53, 1995).

[122] The probes described above may be attached to a solid support, such as paper, membranes, filters, chips, pins, or glass slides, or any other appropriate substrate, such as those further described herein. For example, probes of genes characteristic of small molecule efficacy may be attached covalently or non-covalently to membranes for use, for example, in dotblots, or to solids such as arrays, for example, microarrays.

Drug Design Using Microarrays

[123] The invention also provides methods for designing and optimizing drugs for cancer, for example, those which have been identified as described herein. In one embodiment, compounds may be screened by comparing the expression level of one or more genes characteristic of small molecule efficacy following incubation of a diseased cell of cancer or similar cell with the test compound. In a more preferred embodiment, the expression level of the genes may be determined using microarrays, and comparing the gene expression profile of a cell in response to the test compound with the gene expression profile of a normal cell corresponding to a diseased cell of cancer (a "reference profile"). In a further embodiment, the expression profile may also be compared to that of a diseased cell of cancer. The comparisons are preferably done by introducing the gene expression profile data of the cell treated with drug into a computer system comprising reference gene expression profiles, which are stored in a computer readable form, using appropriate algorithms. Test compounds may be screened for those that alter the level of expression of genes characteristic of small molecule efficacy. Such compounds, that is, compounds which are capable of normalizing the expression of essentially all genes characteristic of small molecule efficacy, are candidate therapeutics.

[124] The efficacy of the compounds may then be tested in additional *in vitro* and *in vivo* assays, and in animal models (e.g., xenograft model). The test compound may be administered to the test animal, and one or more symptoms of the disease may be monitored for improvement of the condition of the animal. Expression of one or more genes characteristic of small molecule efficacy

may also be measured before and after administration of the test compound to the animal. A normalization of the expression of one or more of these genes is indicative of the efficiency of the compound for treating cancer in the animal.

[125] In the clinical setting, obtaining human-derived samples of tissue exhibiting cancer may be difficult, if not prohibitive. Therefore, identification of gene expression changes indicative of efficacy of a therapeutic compound may be determined in a more easily accessible, surrogate cell population, for example, peripheral blood leukocytes (PBLs). This method may be performed either in a human or animal model system. In one embodiment, a test compound may be administered to the test animal (either normal or cancer-containing) at the same doses that have been observed to be efficacious in treating cancer in that animal model. Blood may be drawn from the animal at various time points (e.g., 1, 4, 7, and 24 hours following the first, mid-point, and last day of a regimen of multiple day dosing). Animals dosed with vehicle may be used as controls. RNA may be isolated from PBLs, and can be used to generate probes for hybridization to microarrays. The hybridization results may then be analyzed using computer programs and databases, as described above. The resulting expression profile may be compared directly to the analogous profile from the treated cancer tissue for similarities or simply correlated with efficacy (e.g., in terms of doses and time points) in the animal model.

[126] In another embodiment, human blood may be treated *ex vivo* with a therapeutic compound at a dose consistent with the therapeutic dose in the animal model, or at a dose that is consistent with known plasma levels of the therapeutic dose in the animal model. The blood may be treated (e.g., rocking at 37°C) with the therapeutic compound immediately, or after some period of incubation time (e.g., 24 hours) to allow for gene expression to re-equilibrate after the blood draw. The blood may also be treated with the therapeutic compound for various timepoints (e.g., 4 and 24 hours), and then PBL RNA isolated and used to create a probe for hybridization to a microarray. A compound solubilization agent (e.g., DMSO) may be used as a control. The resulting expression profile may be compared directly to the analogous profile from the treated cancer tissue for similarities or simply correlated with efficacy (e.g., in terms of doses and time points) in the animal model.

[127] The toxicity of the candidate therapeutic compound may be evaluated, for example, by determining whether the compound induces the expression of genes known to be associated with a toxic response. Expression of such toxicity related genes may be determined in different cell types, preferably those that are known to express the genes. In fact, alterations in gene expression may serve as a more sensitive marker of human toxicity than routine preclinical safety studies. In a preferred method, microarrays may be used for detecting changes in the expression of genes known to be associated with a toxic response. It may be possible to perform proof of concept

studies demonstrating that changes in gene expression levels may predict toxic events that were not identified by routine preclinical safety testing (*see, e.g.*, Huang et al., Toxicol. Sci. 63:196-207, 2001; Waring et al., Toxicol. Appl. Pharamacol. 175:28-42, 2001).

Kits

[128] The invention further provides kits for determining the expression level of genes characteristic of small molecule efficacy. The kits may be useful for identifying subjects that are predisposed to developing cancer or who have cancer, as well as for identifying and validating therapeutics for cancer. In one embodiment, the kit comprises a computer readable medium on which is stored one or more gene expression profile of diseased cells of cancer, or at least values representing levels of expression of one or more genes characteristic of small molecule efficacy in a diseased cell. The computer readable medium can also comprise gene expression profiles of counterpart normal cells, diseased cells treated with a drug, and any other gene expression profile described herein. The kit can comprise expression profile analysis software capable of being loaded into the memory of a computer system.

[129] A kit can comprise a microarray comprising probes of genes characteristic of small molecule efficacy. A kit can comprise one or more probes or primers for detecting the expression level of one or more genes characteristic of small molecule efficacy and/or a solid support on which probes attached and which can be used for detecting expression of one or more genes characteristic of small molecule efficacy in a sample. A kit may further comprise nucleic acid controls, buffers, and instructions for use.

[130] Other kits provide compositions for treating cancer. For example, a kit can also comprise one or more nucleic acids corresponding to one or more genes characteristic of small molecule efficacy, e.g., for use in treating a patient having cancer. The nucleic acids can be included in a plasmid or a vector, e.g., a viral vector. Other kits comprise a polypeptide encoded by a gene characteristic of cancer or an antibody to a polypeptide. Yet other kits comprise compounds identified herein as agonists or antagonists of genes characteristic of small molecule efficacy. The compositions may be pharmaceutical compositions comprising a pharmaceutically acceptable excipient.

EXAMPLES

[131] It will be apparent to those skilled in the art that the examples and embodiments described herein are by way of illustration and not of limitation, and that other examples may be used without departing from the spirit and scope of the present invention, as set forth in the claims.

Example 1. Tumor Xenograft Gene Expression Profiling Protocol

A. Tumor implantation and excision

[132] Female nude mice ranging between 11-19 weeks of age, and with an average weight of approximately 18-25 grams were used in these studies. Separately, human colon (HCT-116) cancer cell lines were grown in tissue culture to approximately 70% confluency. Cells were harvested on the day of implant (5×10^6 cells/mouse), and were suspended in Hanks Balanced Salt Solution from the time of harvest to the time of implant at which time each mouse received a 0.2 ml injection of the cell suspension of the appropriate cell innoculum. The cells were injected subcutaneously in the right flank of each mouse and tumors were monitored for growth. Time of staging (dosing) was determined when tumors reached a size of 75-125 mgs (from Day 5 - Day 9 of implant). The vehicle used for the histone deacetylase inhibitor is 12.5% ethanol, 12.5% Cremafor EL, water, or saline. For the HCT-116 xenografts, dosing was intravenous at 25 mg/kg once a day for 3 days. On day 3, at 3, 6, and 24 hours after the dose, mice are euthanized and two drug-treated and two vehicle-treated tumors are harvested and snap frozen. Table 1 describes the genes that are up- and down-regulated greater than or equal to two-fold in the histone deacetylase inhibitor-treated HCT-116 cells relative to vehicle-treated cells.

B. RNA extraction and cRNA preparation

[133] Total RNA was extracted from tumor explants or cell lines using TRIzol reagent (Life Technologies, MD) according to a modified vendor protocol which utilizes the RNeasy protocol (Qiagen, CA). After homogenization with a Brinkmann Polytron PT10/35 (Brinkmann, Switzerland) and phase separation with chloroform, samples were applied to RNeasy columns. RNA samples were treated with DNase I using Rnase-free DNase Set (Qiagen, CA).

[134] After elution and quantitation with UV spectrophotometry, each sample was reverse transcribed into double-stranded cDNA using the Gibco Superscript II Choice System for RT-PCR according to vendor protocol (Invitrogen, CA).

[135] Samples were organically extracted and ethanol precipitated. Approximately 1 μg cDNA was then used in an *in vitro* transcription reaction incorporating biotinylated nucleotides using an RNA labeling kit (Enzo Diagnostics, NY). The resulting cRNA was then subjected to RNeasy clean-up protocol and then quantified using UV spectrophotometry. The cRNA (15 μg) was fragmented in the presence of MgOAc and KOAc at 94°C. Fragmented RNA (10 μg) was loaded

onto each array, one cRNA sample per array. Arrays were hybridized for 16 hours at 45°C rotating at 60 rpm in an Affymetrix GeneChip Hybridization Oven 640.

C. Microarray Suite 5.0 analysis

[136] Following hybridization, arrays were stained with Phycoerythrin-conjugated Streptavidin, placed in an Agilent GeneArray Scanner and then exposed to a 488 nm laser, causing excitation of the phycoerythrin. The Microarray Suite 5.0 software digitally converts the intensity of light given off by the array into a numeric value indicative of levels of gene expression. Because each array represented a single animal sample, treated animals were compared to the vehicle animals and relative fold changes of genes were obtained. Those genes increased or decreased by at least 2-fold change were considered significant and chosen for further analysis.

Table 1

Genes up- and down-regulated by a histone deacetylase inhibitor in colon HCT-116 xenografts.

Probe Set	Gene Symbol	Title	RMA Rank	3 hr Fold Change	6 hr Fold Change	24 hr Fold Change	Genebank ID
200696_s_at	GSN	"gelsolin (amyloidosis, Finnish type)"	1257	1.28	1.32	1.54	NM_000177
202033_s_at	RB1CC1	RB1-inducible coiled-coil 1	2	3.55	4.57	2.56	BG402105
202284_s_at	CDKN1A	"cyclin-dependent kinase inhibitor 1A (p21, Cip1)"	11243	1.07	1.05	1.07	NM_000389
202589_at	TYMS	thymidylate synthetase	-14	-1.77	-2.19	-1.39	NM_001071
202954_at	UBE2C	ubiquitin-conjugating enzyme E2C	-11	-2.44	-2.49	-1.79	NM_007019
203708_at	PDE4B	"phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)"	-1	-7.43	-4.68	-0.87	NM_002600
204268_at	S100A2	S100 calcium binding protein A2	-2	-2.05	-2.96	-2.07	NM_005978
204420_at	FOSL1 (FRA-1)	FOS-like antigen 1	-18	-2.16	-2.93	-1.97	BG251266
205352_at	SERPINI1	"serine (or cysteine) proteinase inhibitor, clade I (neuroserpin), member 1"	11	3.90	6.54	2.60	NM_005025
206463_s_at	DHRS2	dehydrogenase/reductase (SDR family) member 2	6	3.60	7.13	3.28	NM_005794
209727_at	GM2A	GM2 ganglioside activator protein	3806	3.10	2.17	2.40	M76477
211698_at	CRI1	CREBBP/EP300 inhibitory protein 1	830	2.04	2.36	2.22	AF349444
212464_s_at	FN1	fibronectin 1	487	3.67	6.12	9.08	X02761
214079_at	---	"Homo sapiens cDNA FLJ20338 fis, clone HEP12179, mRNA sequence"	1	3.47	6.58	2.64	AK000345
214710_s_at	CCNB1	cyclin B1	-16	-2.21	-2.51	-2.27	BE407516
216060_s_at	DAAM1	dishevelled associated activator of morphogenesis 1	8	3.16	2.61	0.98	AK021890
217761_at	SIPL	SIPL protein	-24	-2.01	-2.37	-1.68	NM_018269
202094_at	BIRC5	baculoviral IAP repeat-containing 5 (survivin)		-1.52	-1.35	-0.02	NM_001168

ABSTRACT

The present invention relates to gene expression profiles, microarrays comprising nucleic acid sequences representing gene expression profiles, and methods of using gene expression profiles and microarrays. The invention also provides methods and compositions for diagnostic assays for detecting cancer and therapeutic methods and compositions for treating cancer. The invention also provides methods for designing, identifying, and optimizing therapeutics for cancer. Diagnostic compositions of the invention include compositions comprising detection agents for detecting one or more genes that have been shown to be up- or down-regulated in cells of cancer treated with therapeutics (compounds) relative to untreated counterpart cells. Exemplary detection agents include nucleic acid probes, which can be in solution or attached to a solid surface, e.g., in the form of a microarray. The invention also provides computer-readable media comprising values of levels of expression of one or more genes that are up- or down-regulated in compound- or therapeutic-treated cancer cells.

Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/US04/041883

International filing date: 10 December 2004 (10.12.2004)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/529,432
Filing date: 12 December 2003 (12.12.2003)

Date of receipt at the International Bureau: 26 January 2005 (26.01.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse